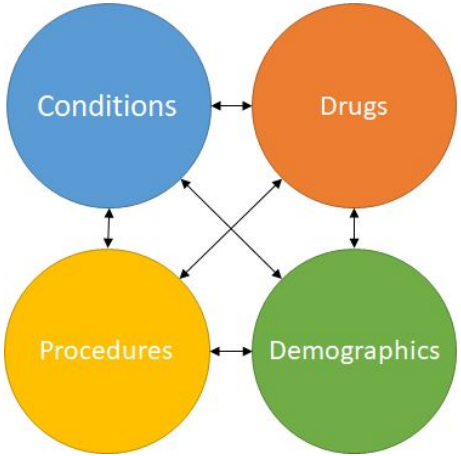
Clinical Data Services Provider

Translator Kick-Off Relay
April 27, 2020

COHD: Columbia Open Health Data

Columbia Open Health Data (COHD)

- Clinical associations mined from observational EHR data
- Conditions, drugs, procedures, gender, race, ethnicity
- EHR prevalence: # patients with each concept
- Co-occurrence count: # patients with each pair of concepts
- Associations calculated from EHR prevalence and co-occurrence count
- Privacy protection measures
  - Exclude counts ≤ 10
  - Perturb counts using Poisson distribution
- http://cohd.io/api



	Count
Patients	5 M
Condition concepts	10 K
Drug concepts	10 K
Procedure concepts	10 K
Pairs of concepts	50 M

COHD Association Analyses

1) Relative Frequency (conditional probability): $F_R(C_1|C_2) = \frac{N_{C_1,C_2}}{N_{C_2}}$

   Example:

   F_R(Atrial Fibrillation | Warfarin) = 57.4%

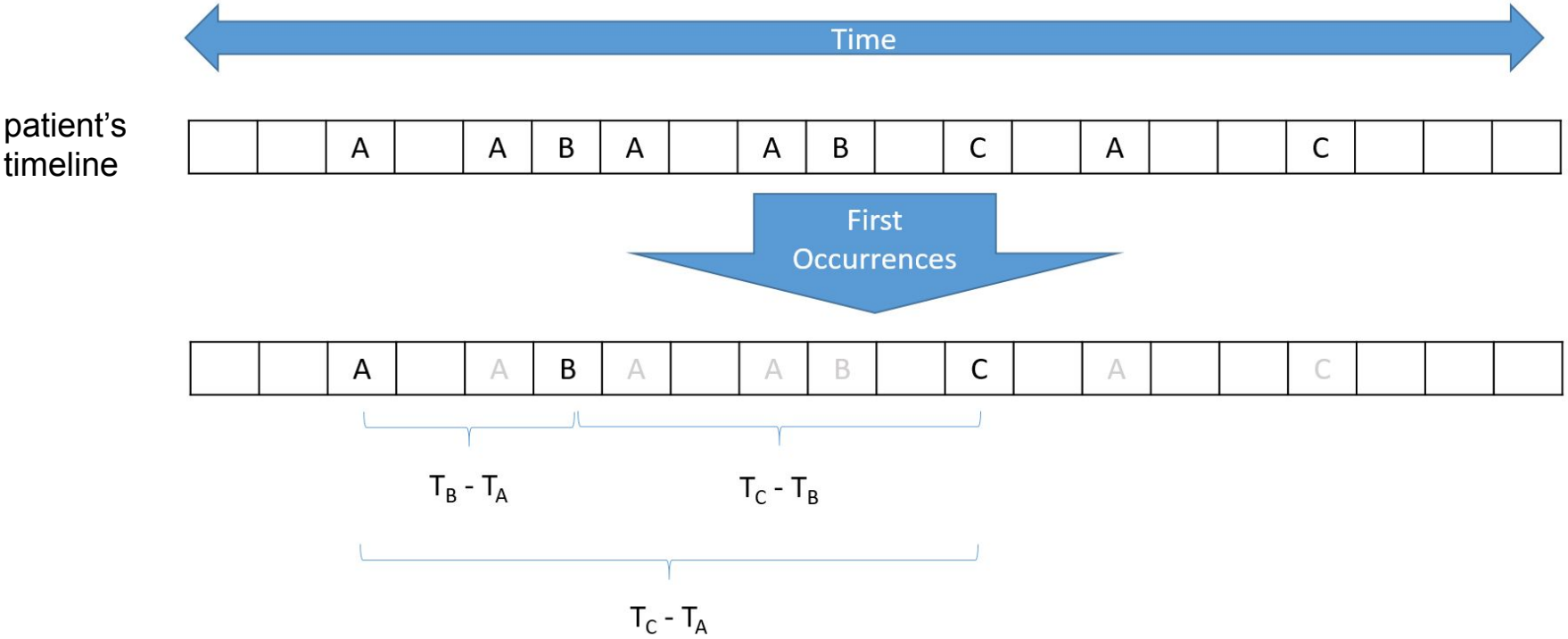
2) Observed-Expected Frequency Ratio (association strength):

   $$LR(C_1, C_2) = \log_e \frac{N_{C_1,C_2} \cdot N_P}{N_{C_1} \cdot N_{C_2}}$$

   Example:

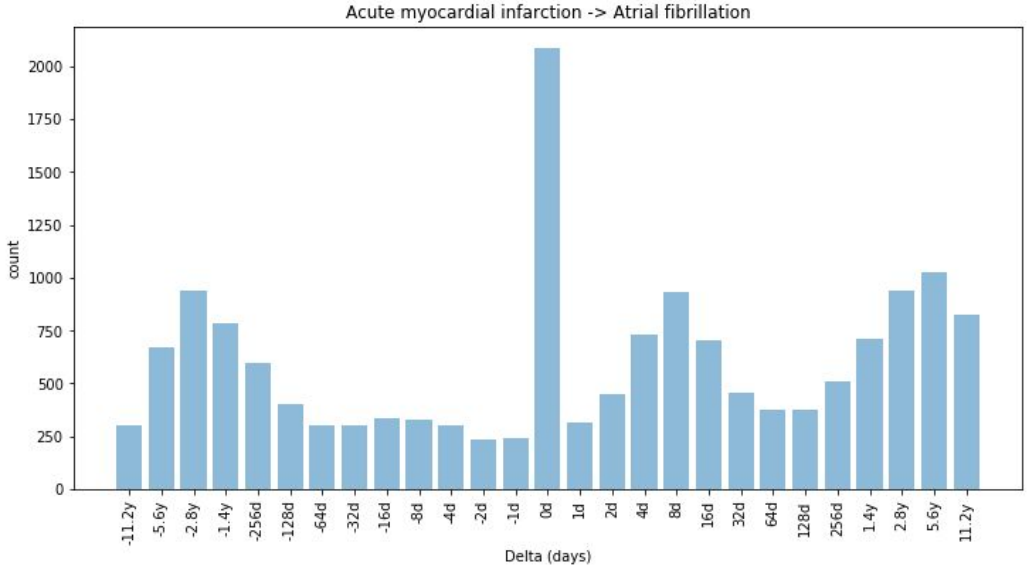
   LR(Atrial Fibrillation , Warfarin) = 2.93

COHD: Temporal Co-occurrences (alpha)



Delta: time difference between first occurrences of a pair of concepts

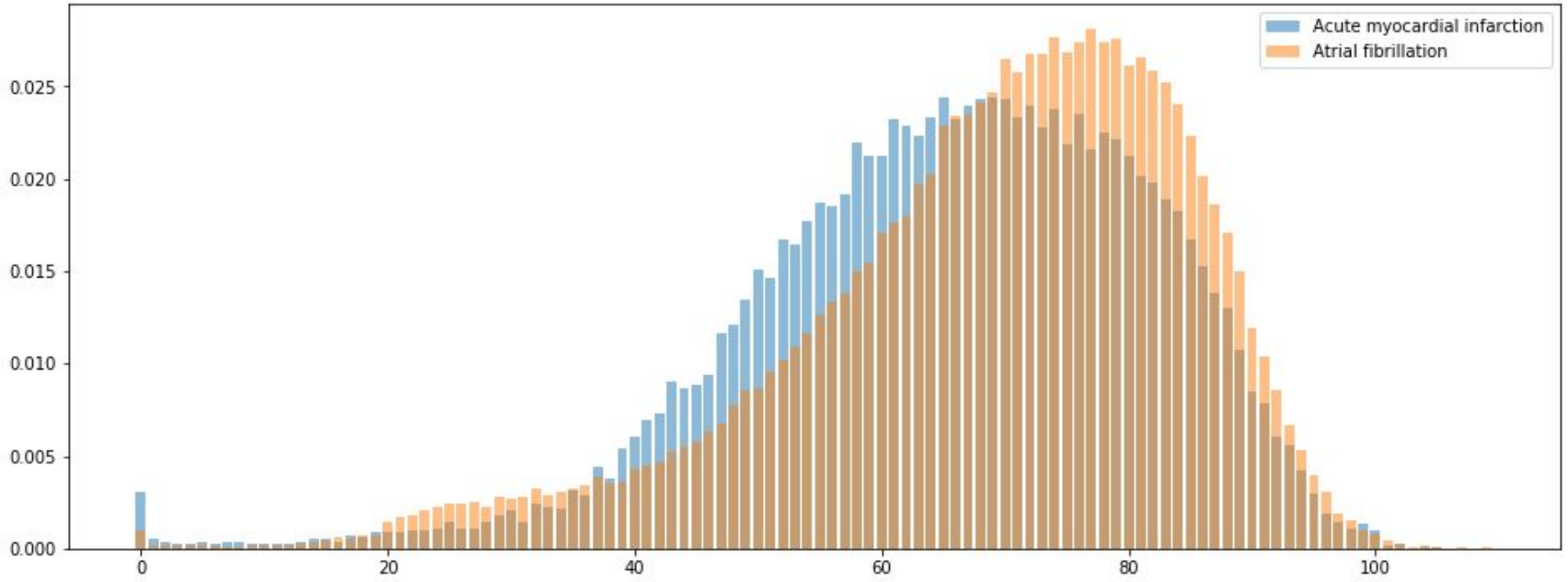
Delta Distributions



Delta (days)	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Bin	3				2		-1	0	1	2		3			

- Variable bin sizes for counting deltas
- Bin widths grow by powers of 2 (0 days, 1 days, 2 days, 4 days, 8 days, …)

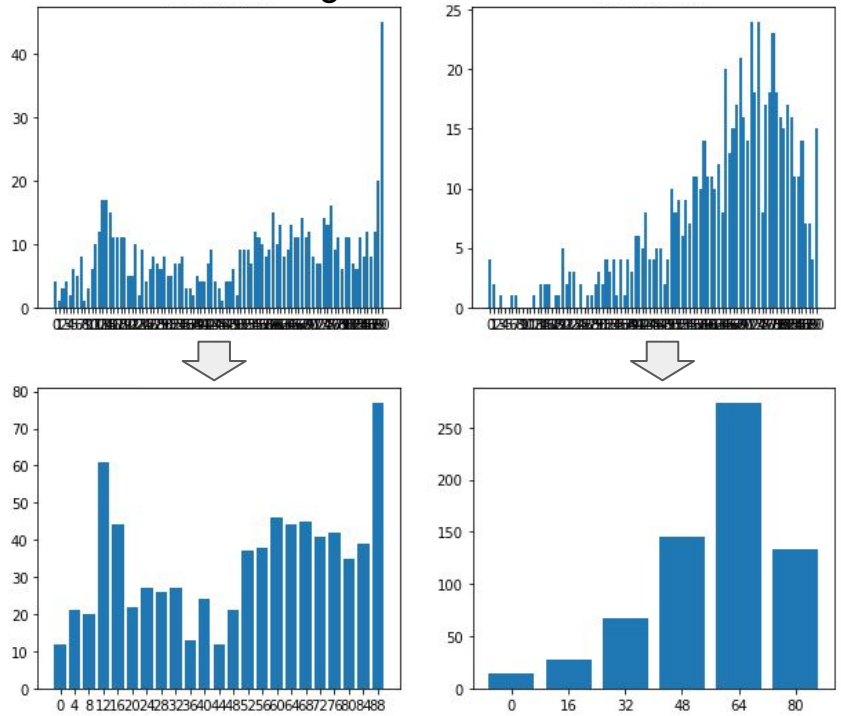
Concept-Age Distributions



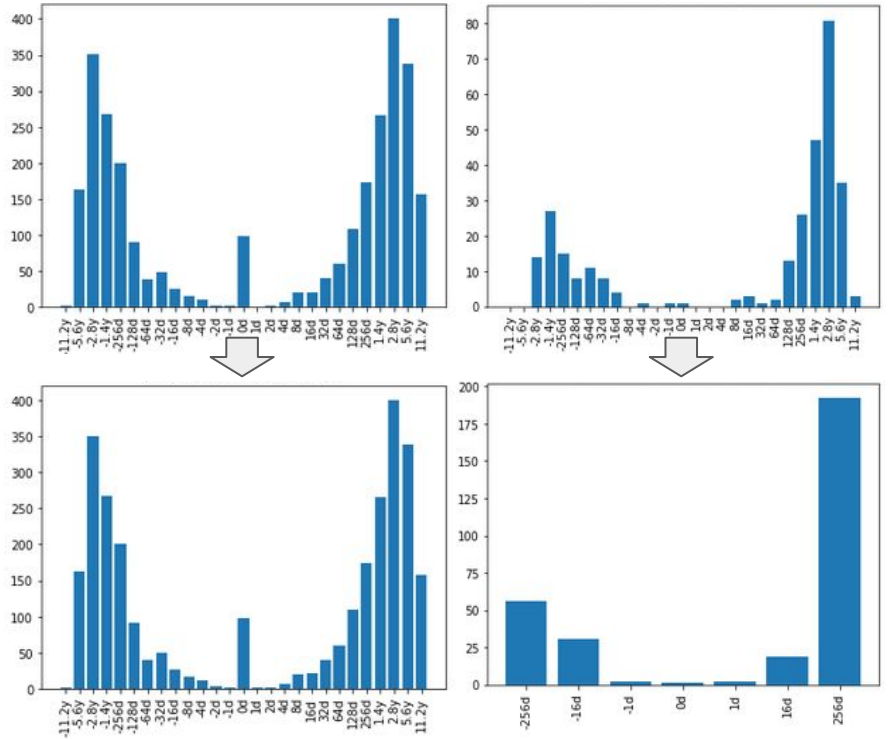
- Calculate the age-distributions for each concept to account for expected differences in age of occurrence

Patient Privacy Protections

Age Distributions



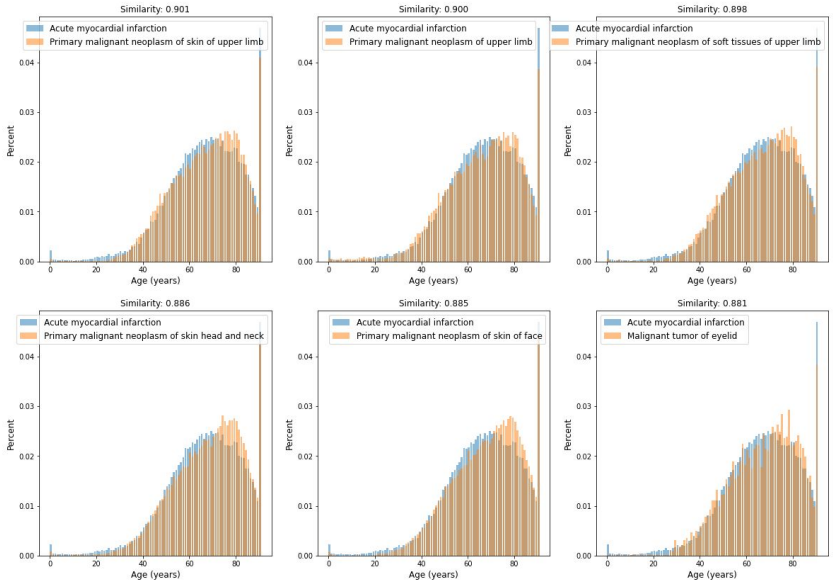
Delta Distributions



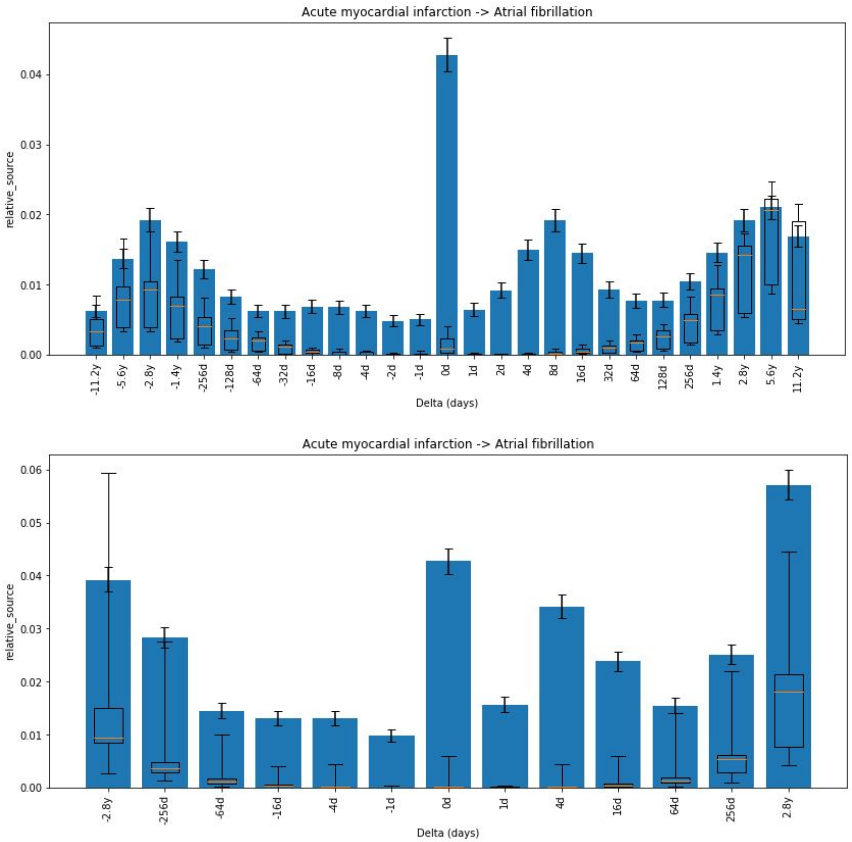
- Suppress counts < 10
- Poisson perturbation
- Adaptive binning of age and delta distributions

Comparison Distribution (WIP)

Similar Concepts



Comparison Distributions

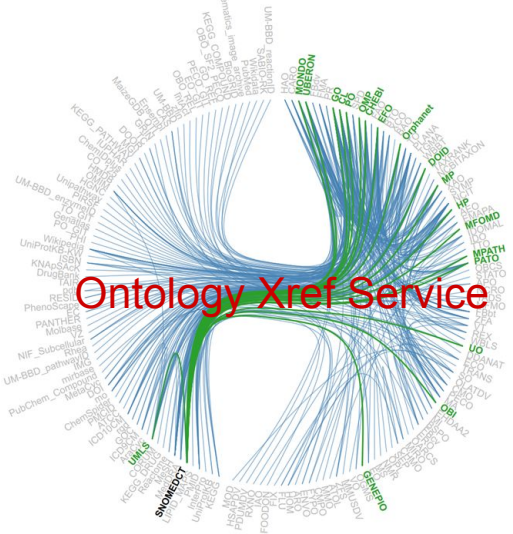


Find similar concepts based on age-distribution to create comparison distributions

Concept Normalization

OMOP standard concepts

{SNOMED, ICD9, ICD10, MeSH, UMLS}



Query Graph

```
"query_graph": {
  "nodes": [
    {
      "id": "n00",
      "curie": "DOID:0060224",
      "type": "condition"
    },
    {
      "id": "n01",
      "curie": "CHEBI:10033"
    }
  ],
  "edges": [
    {
      "id": "e00",
      "type": "association",
      "source_id": "n00",
      "target_id": "n01"
    }
  ]
}
```

Response

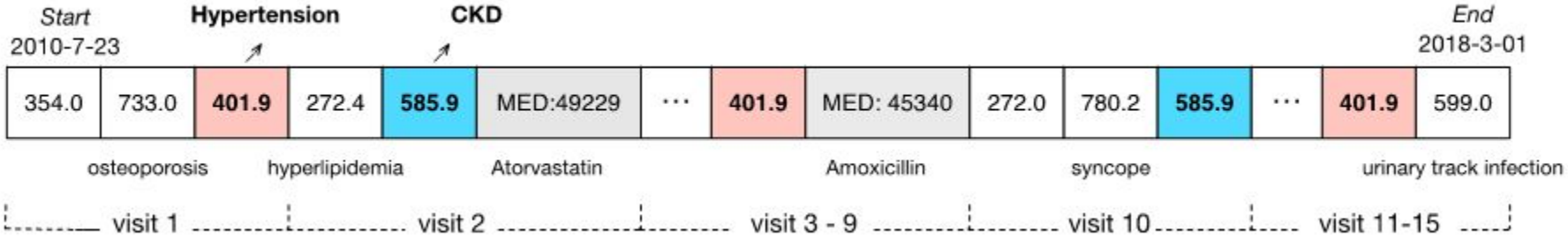
```
"nodes": [
    {
      "id": "OMOP:313217",
      "name": "Atrial fibrillation",
      "omop_domain": "Condition",
      "synonyms": [
        {
          "distance": 1,
          "target_curie": "DOID:0060224",
          "target_label": "atrial fibrillation"
        }
      ],
      "type": "disease"
    },
    {
      "id": "OMOP:1310149",
      "name": "Warfarin",
      "omop_domain": "Drug",
      "synonyms": [
        {
          "distance": 2,
          "target_curie": "CHEBI:10033",
          "target_label": "warfarin"
        }
      ],
      "type": "drug"
    }
  ]
```

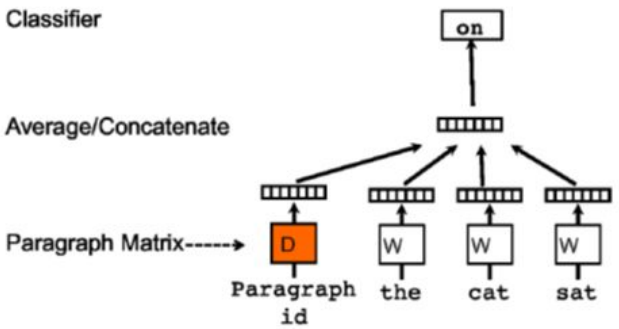
Architecture Challenges

- Normalization of highly specific clinical concepts, e.g.,
  “Warfarin Sodium 2.5 MG Oral Tablet [Coumadin]” (RxNorm:855314)
  “Ultrasonic guidance for needle placement (eg, biopsy, aspiration, injection, localization device), imaging supervision and interpretation” (CPT4:76942)
- Edges representing different association metrics (relative frequency, association strength, temporal association)
- Representing time context of temporal associations

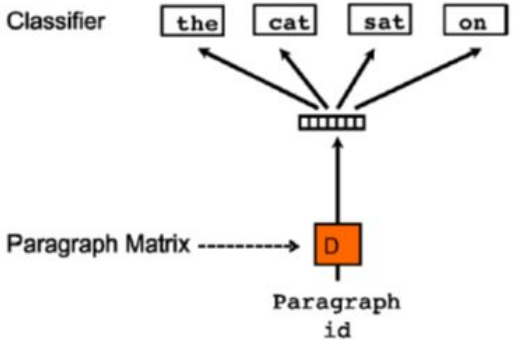
Disease Subtyping with Patient Vectors

Patient Vector Embedding





Distributed Memory (DM)



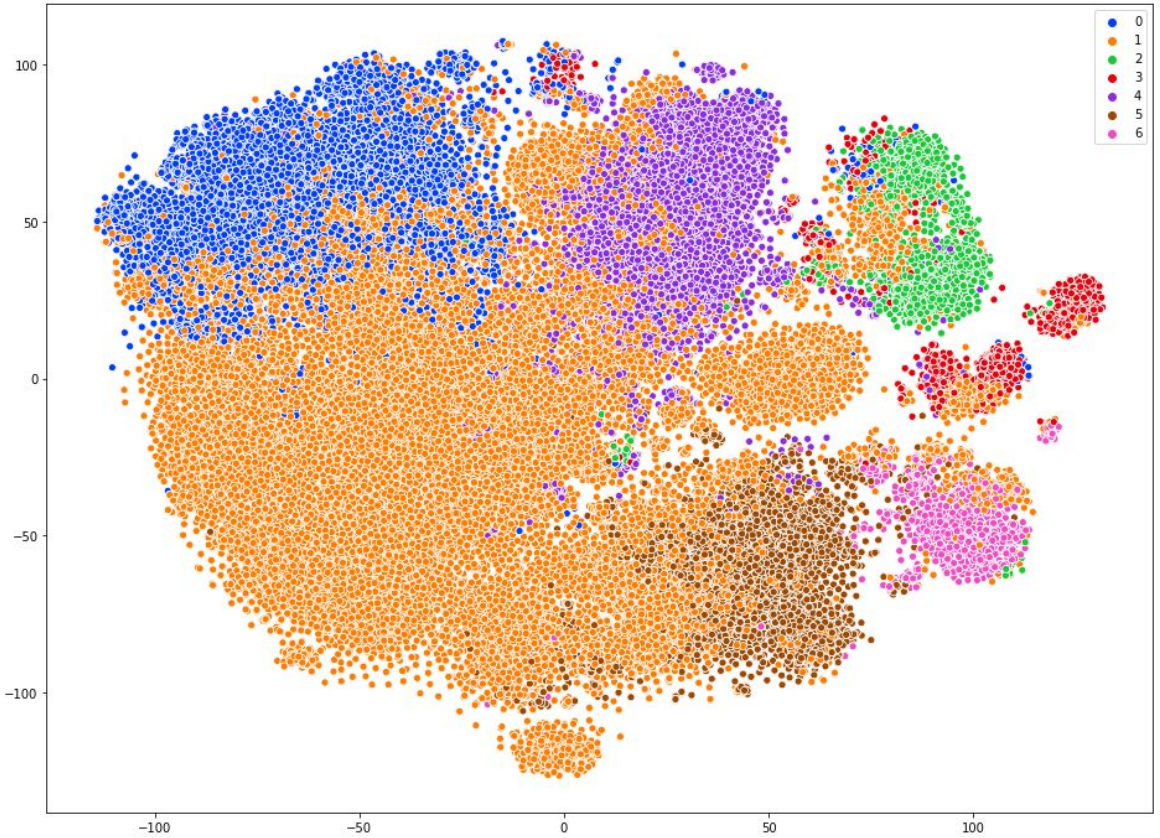
Distributed Bag of Words (DBOW)

- Patient medical sequence: conditions, drugs, procedures, labs sequenced temporally from entire patient timeline
- Apply Doc2Vec (DM + DBOW models) to train patient vector embeddings (200-dim)

Disease Subtyping

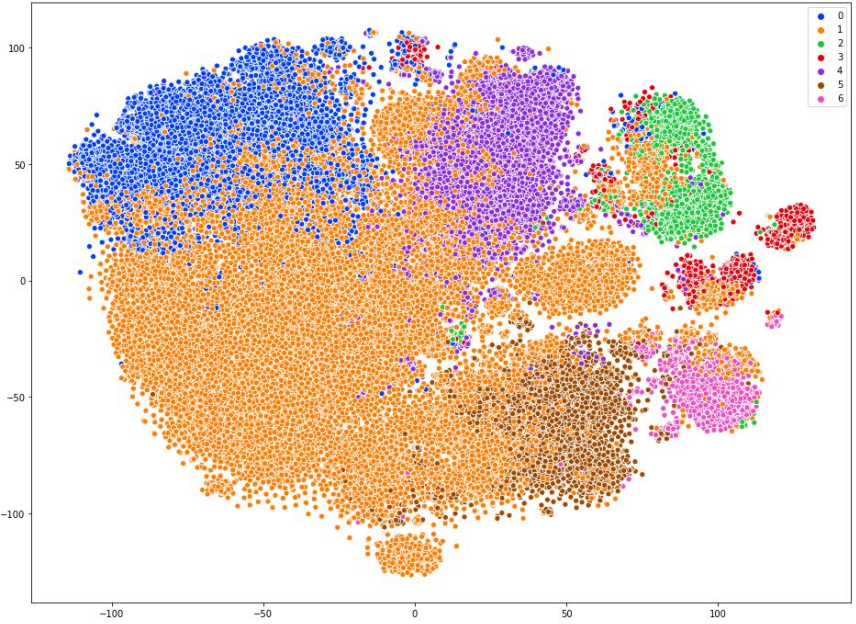
- Sample domain: chronic kidney disease (CKD)
- Automatic selection of disease subtypes using OMOP concept hierarchy
  - Chronic kidney disease stage 1
  - Chronic kidney disease stage 2
  - Chronic kidney disease stage 3
  - Chronic kidney disease stage 4
  - Chronic kidney disease stage 5
  - Chronic kidney disease due to type 2 diabetes mellitus
  - Chronic kidney disease due to hypertension
    - Excluded as an iatrogenic code
- Extract 1-year time window (±180 days) around first occurrence of each condition
- Patient Vector inference

K-Means Clustering



- Learn disease subtypes using K-Means clustering
- Try to find new subtypes → K = # known subtypes + 1

Learned (Clustered) Subtype: #2 vs #0



#0: T2DM, heart conditions (Congestive heart failure, Heart failure, Chest pain, Atrial fibrillation)
#2: End-stage renal disease, Transplanted kidney, Anemia

concept_name	count	prevalence
Type 2 diabetes mellitus	6577	75.182899
Chronic kidney disease due to type 2 diabetes ...	6295	71.959305
Acute renal failure syndrome	6240	71.330590
Hyperlipidemia	6123	69.993141
Essential hypertension	6108	69.821674
Chronic kidney disease	5888	67.306813
Congestive heart failure	5822	66.552355
Atherosclerosis of coronary artery without ang...	5544	63.374486
Dyspnea	5293	60.505258
Abnormal findings on diagnostic imaging of lung	5018	57.361683
Atelectasis	4928	56.332876
Type 2 diabetes mellitus without complication	4853	55.475537
Disorder of body system	4710	53.840878
Pleural effusion	4633	52.960677
Chronic kidney disease stage 3	4626	52.880658
Heart failure	4577	52.320530
Chronic pulmonary edema	3811	43.564243
Chest pain	3524	40.283493
Atrial fibrillation	3456	39.506173
Anemia	3261	37.277092

concept_name	count	prevalence
Essential hypertension	2442	88.542422
End-stage renal disease	2430	88.107324
Anemia in chronic kidney disease	2233	80.964467
Acute renal failure syndrome	1860	67.440174
Chronic kidney disease stage 3	1805	65.445975
Transplanted kidney present	1789	64.865845
Anemia	1778	64.467005
Chronic kidney disease	1642	59.535896
Chronic kidney disease stage 4	1634	59.245830
Chronic kidney disease stage 5	1485	53.843365
Acute tubular necrosis	1362	49.383611
Benign essential hypertension	1348	48.875997
Hyperkalemia	1254	45.467730
Hyperlipidemia	1252	45.395214
Atelectasis	890	32.269761
Urinary tract infectious disease	880	31.907179
Dyspnea	829	30.058013
Abnormal chest sounds	816	29.586657
Type 2 diabetes mellitus	784	28.426396
Acute posthemorrhagic anemia	774	28.063814

Data2Services

A framework and Command Line Interface for building and deploying Translator data and services in a reproducible manner.

Problems

- Various tools and interfaces are required to **realize the vision of a Translator ecosystem**
  - RDF / property graphs
  - BioThings APIs, Reasoner APIs
  - Data providers (e.g. Docket)
- Data providers could benefit from additional guidance to **expose their structured data** with **Translator-compliant interfaces**
- Transformation workflows are usually implemented per case and can be **hard to reconfigure**

Data2Services

Test on local machine | Deploy on single server | Scale in cluster

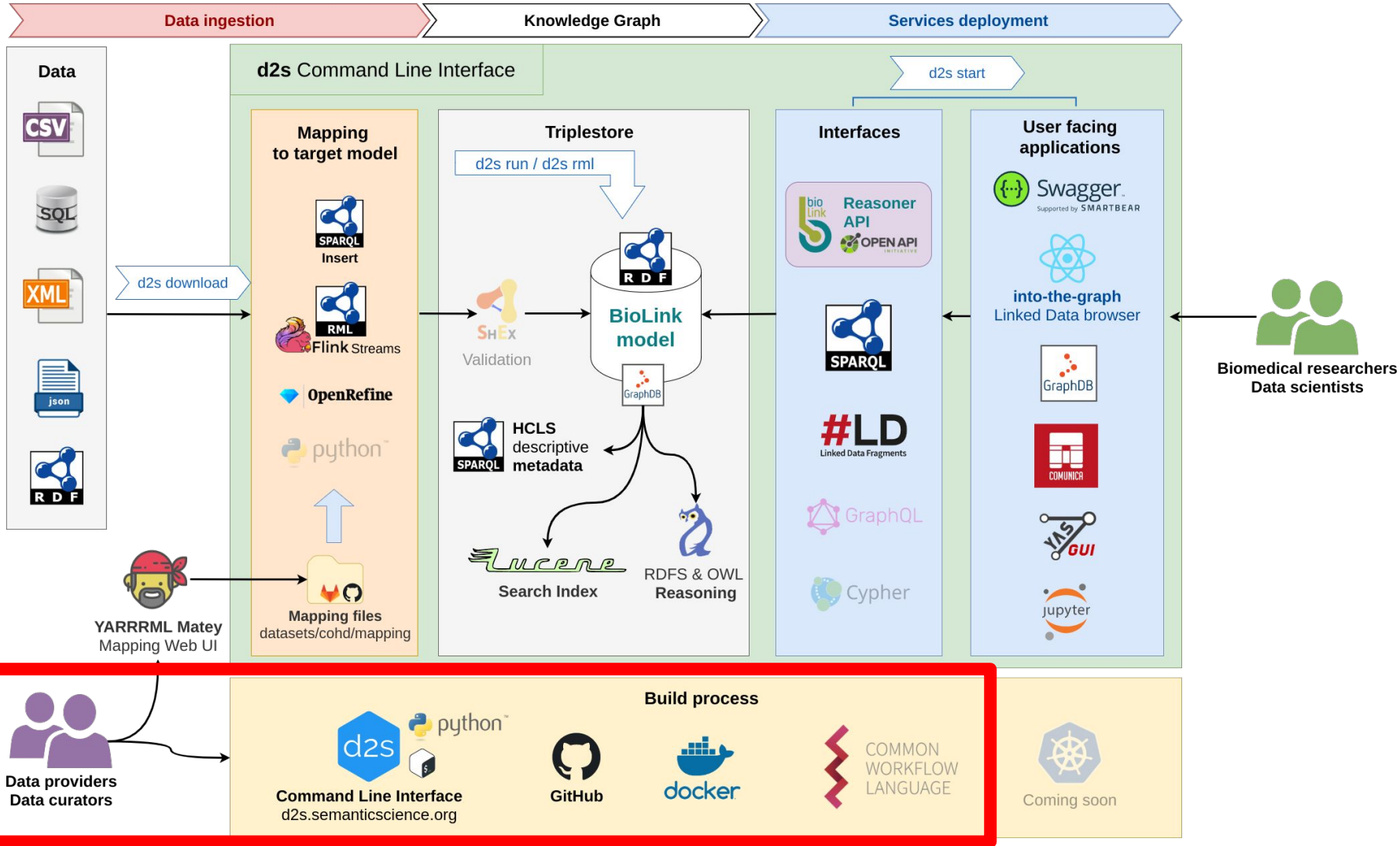
Container-based deployment of **services and workflows** on a Linux or MacOS laptop 💻

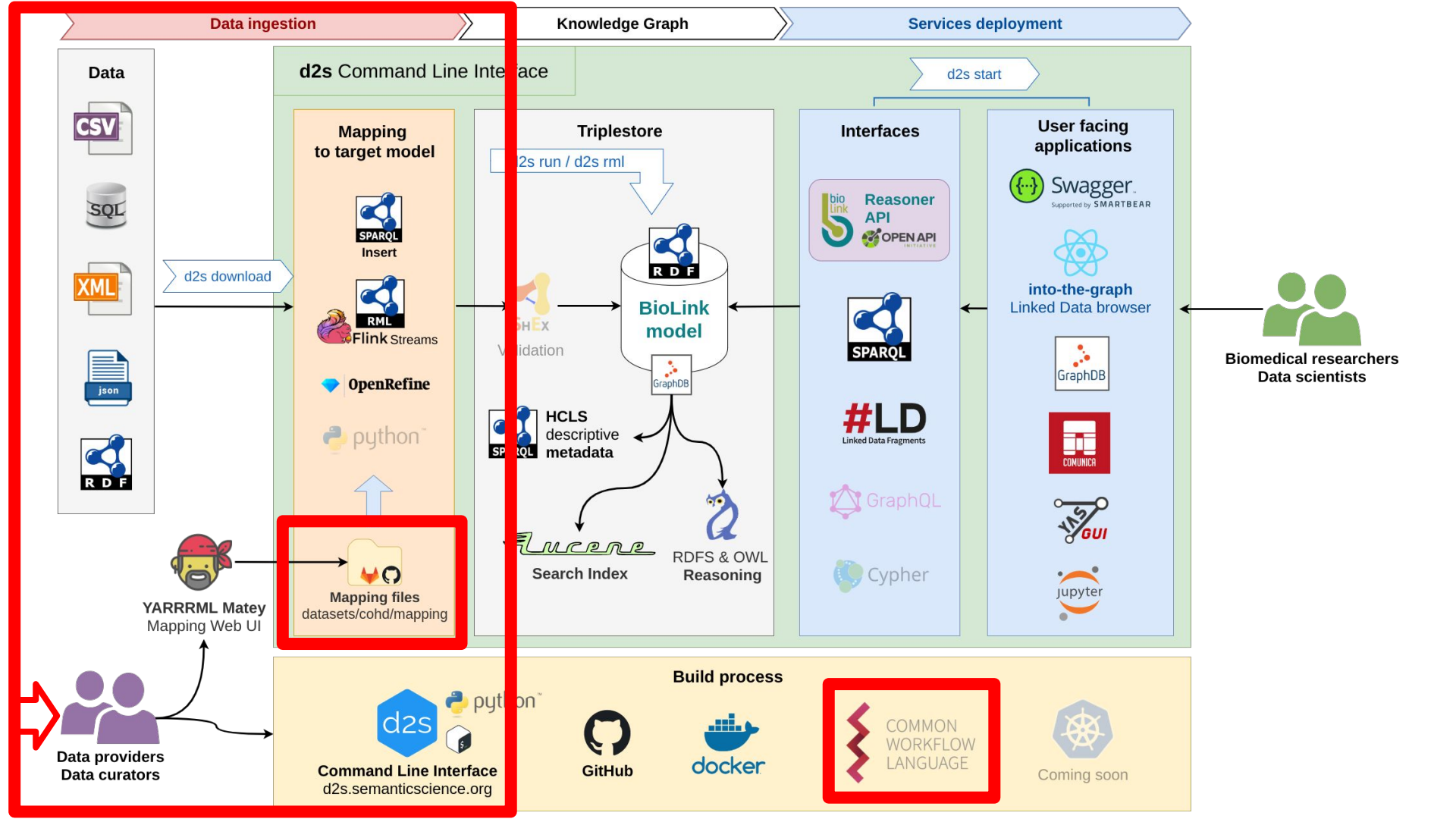
Container-based deployment of **services and workflows** on a single Linux server

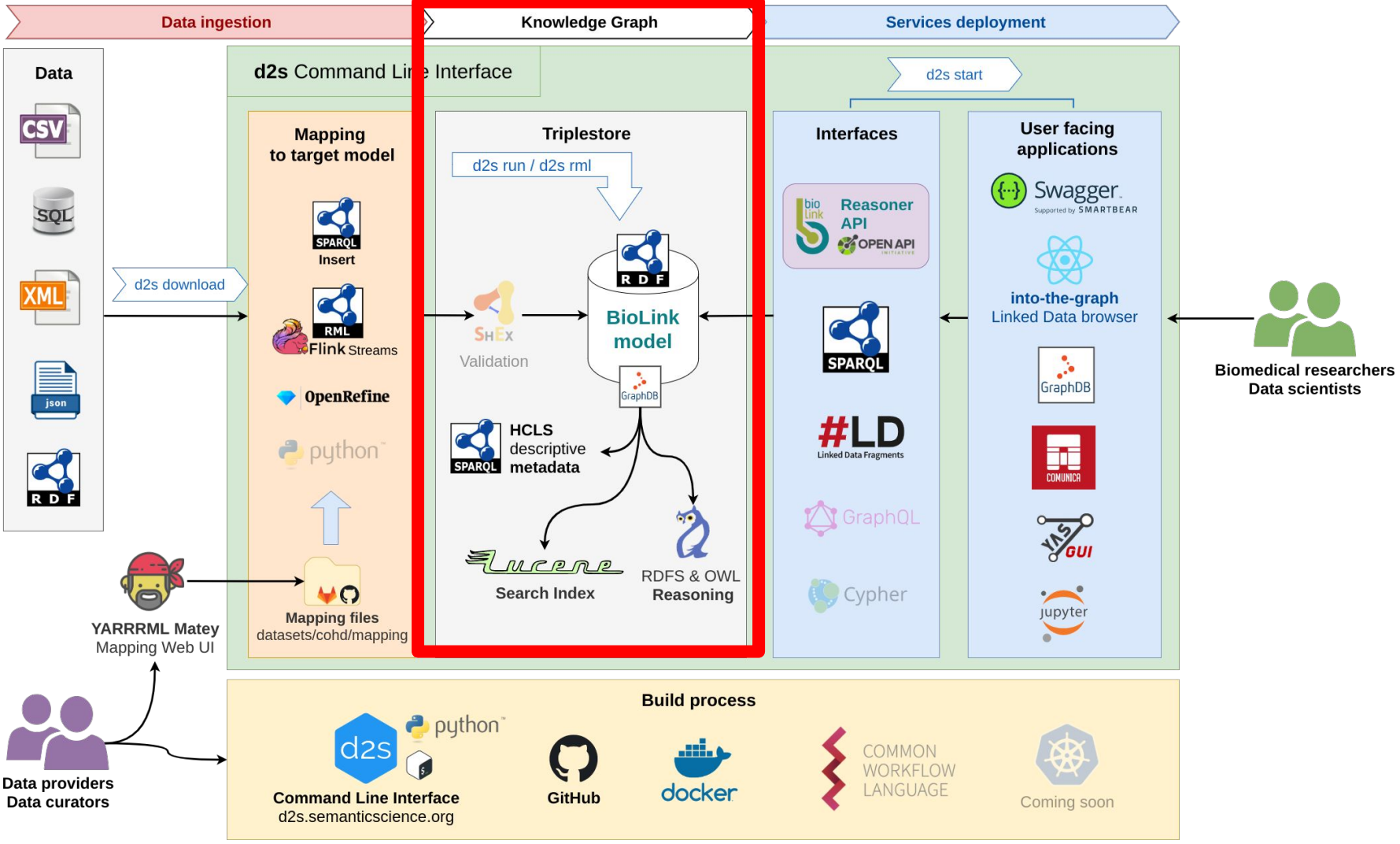
In development: deploy on multiple nodes in a cluster with Kubernetes or OpenShift

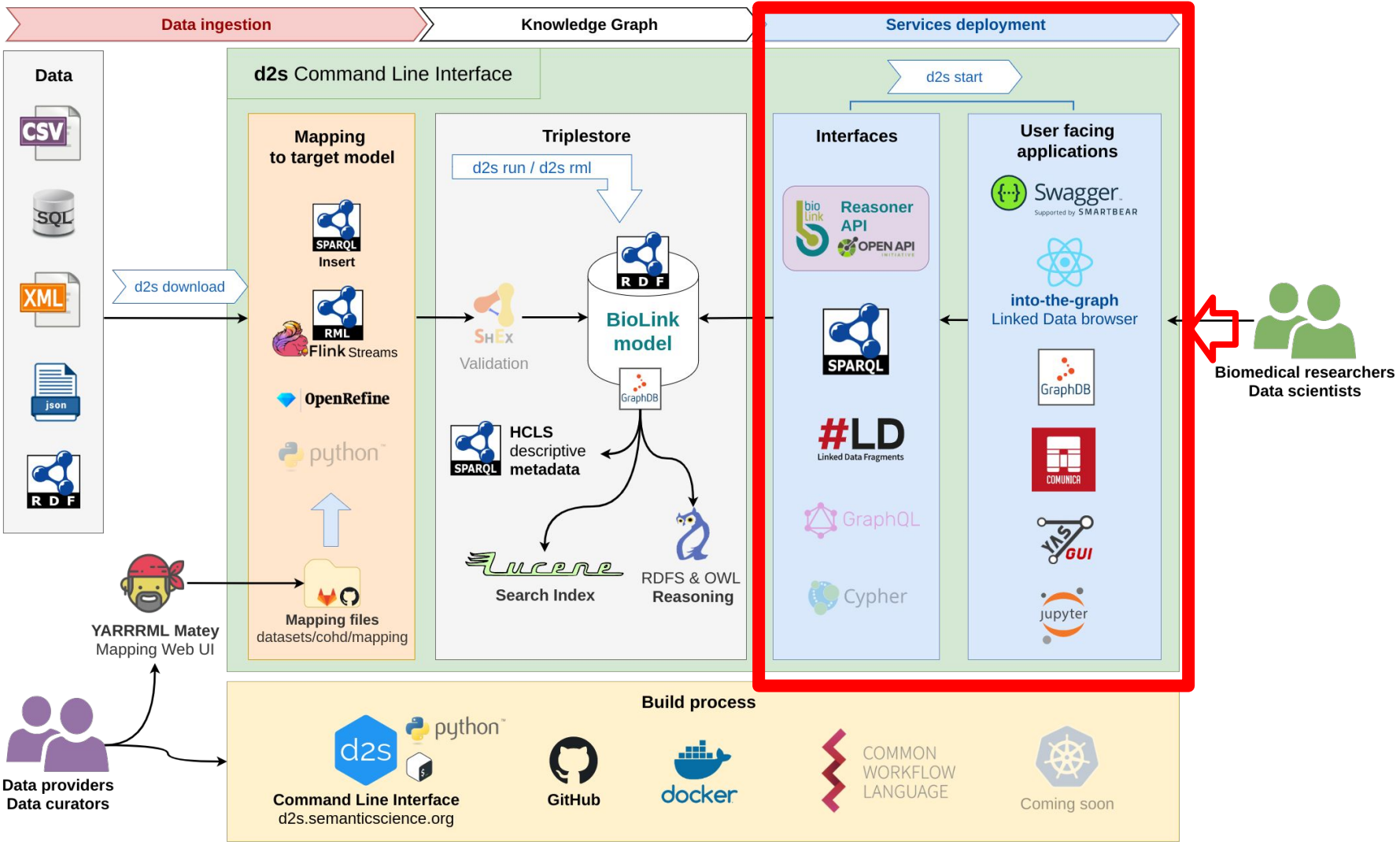
Tested on

Work in progress: 


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s start
d2s download
Mapping to target model
SPARQL Insert
RML Flink Streams
OpenRefine
python
Triplestore
d2s run / d2s rml
SHEX Validation
RDF
BioLink model
GraphDB
SPARQL HCLS descriptive metadata
Lucene Search Index
RDFS & OWL Reasoning
Interfaces
bio link Reasoner API
OPEN API
SPARQL
#LD Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger Supported by SMARTBEAR
into-the-graph Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
YARRRML Matey
Mapping Web UI
Mapping files
datasets/cohd/mapping
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s start
Mapping to target model
SPARQL
Insert
RML
Flink Streams
OpenRefine
python
d2s download
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Lucene
Search Index
RDFS & OWL Reasoning
Interfaces
bio link
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YAS GUI
jupyter
Biomedical researchers
Data scientists
YARRRML Matey
Mapping Web UI
Mapping files
datasets/cohd/mapping
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon

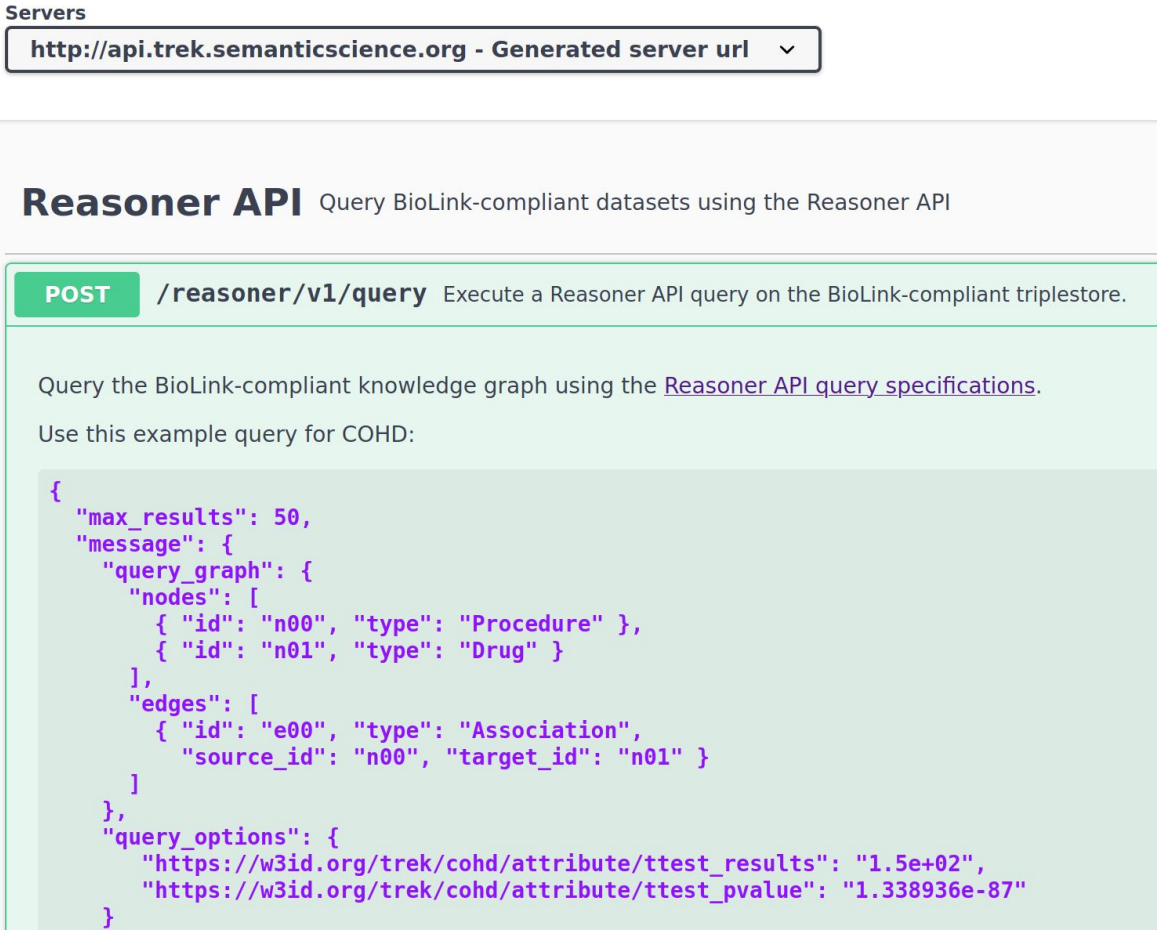

Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s start
Mapping to target model
SPARQL Insert
RML Flink Streams
OpenRefine
python
d2s download
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Lucene
Search Index
RDFS & OWL Reasoning
Interfaces
bio link
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YAS GUI
jupyter
Biomedical researchers
Data scientists
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s download
Mapping to target model
SPARQL Insert
RML Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX Validation
RDF
BioLink model
GraphDB
SPARQL HCLS descriptive metadata
Lucene
Search Index
RDFS & OWL Reasoning
d2s start
Interfaces
bio link Reasoner API
OPEN API
SPARQL
#LD Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YAS GUI
jupyter
Biomedical researchers
Data scientists
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon

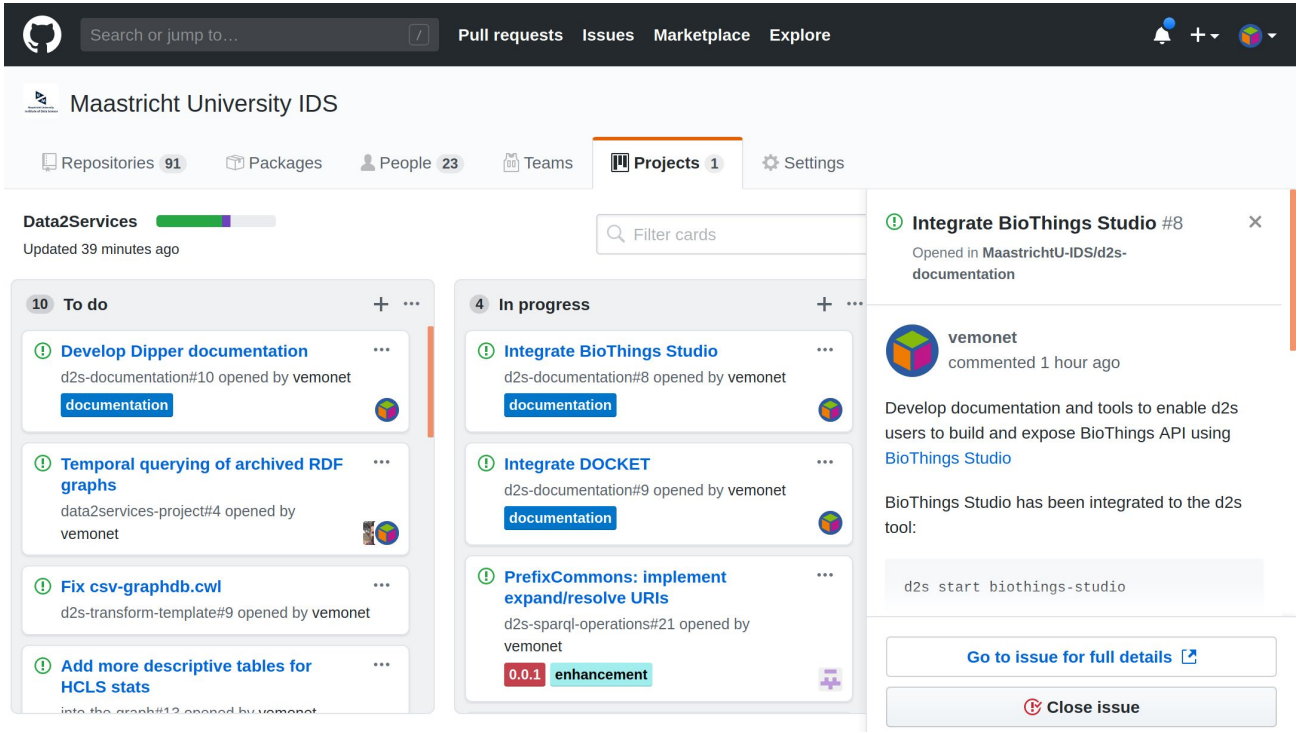
Query BioLink RDF through a ReasonerStd API

https://github.com/MaastrichtU-IDS/d2s-api

Built using Spring Boot framework



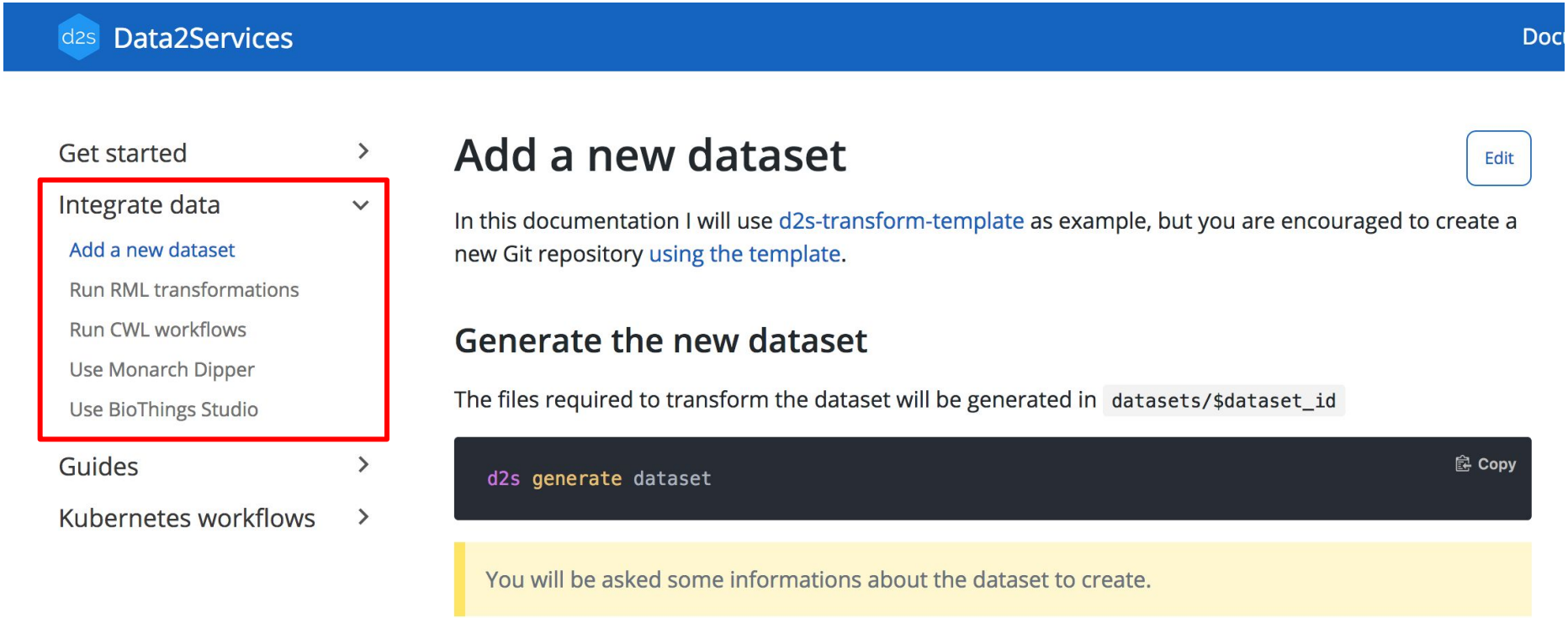
Integrating the Translator solutions



⚠️ GitHub issues board: https://github.com/orgs/MaastrichtU-IDS/projects/3

💬 Discussions and pull requests welcome! We have Gitter: um-dsri/data2services

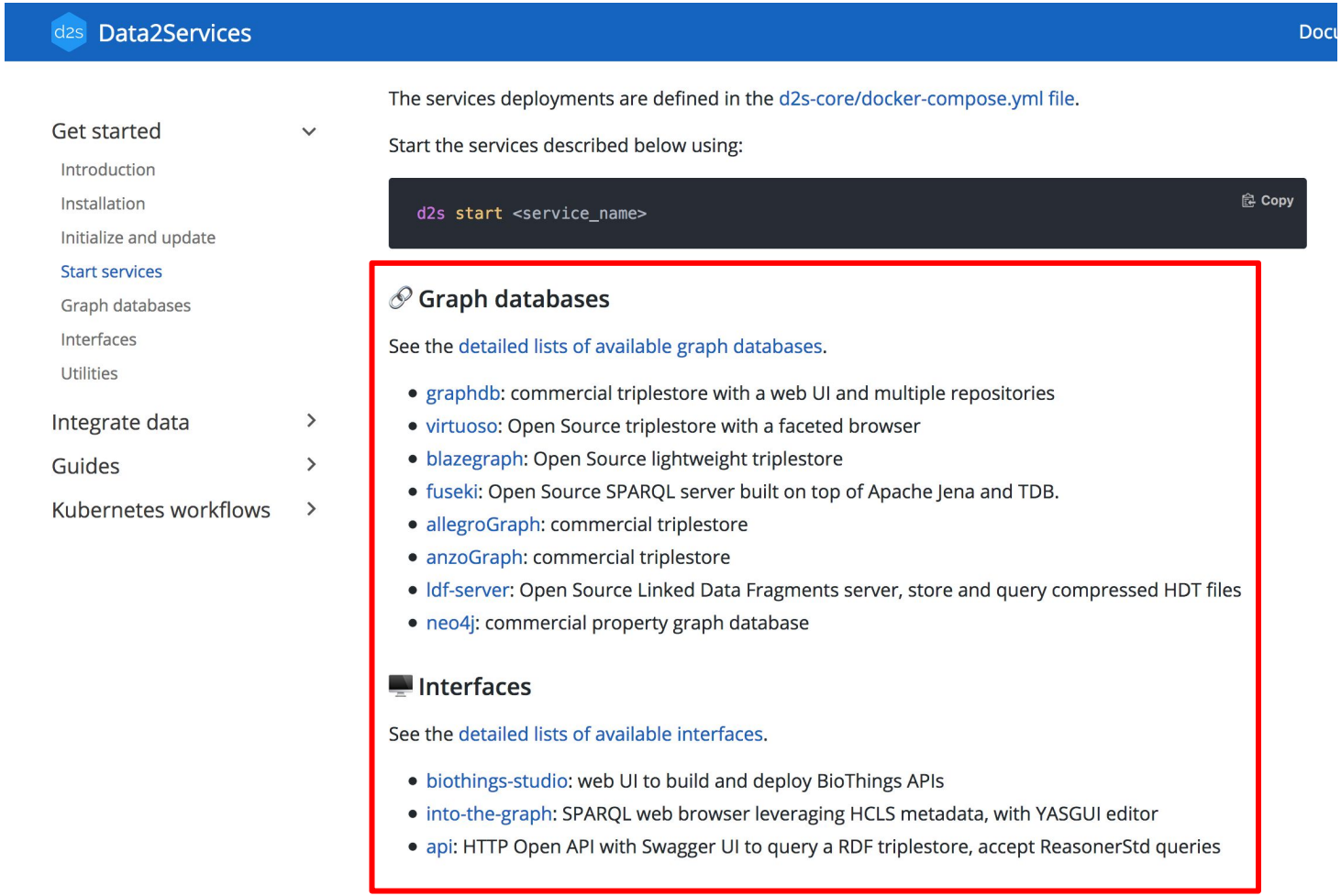
Guide Translator users through multiple solutions to integrate their data



🗃️ Use common tools and project structure (metadata, mappings, scripts)

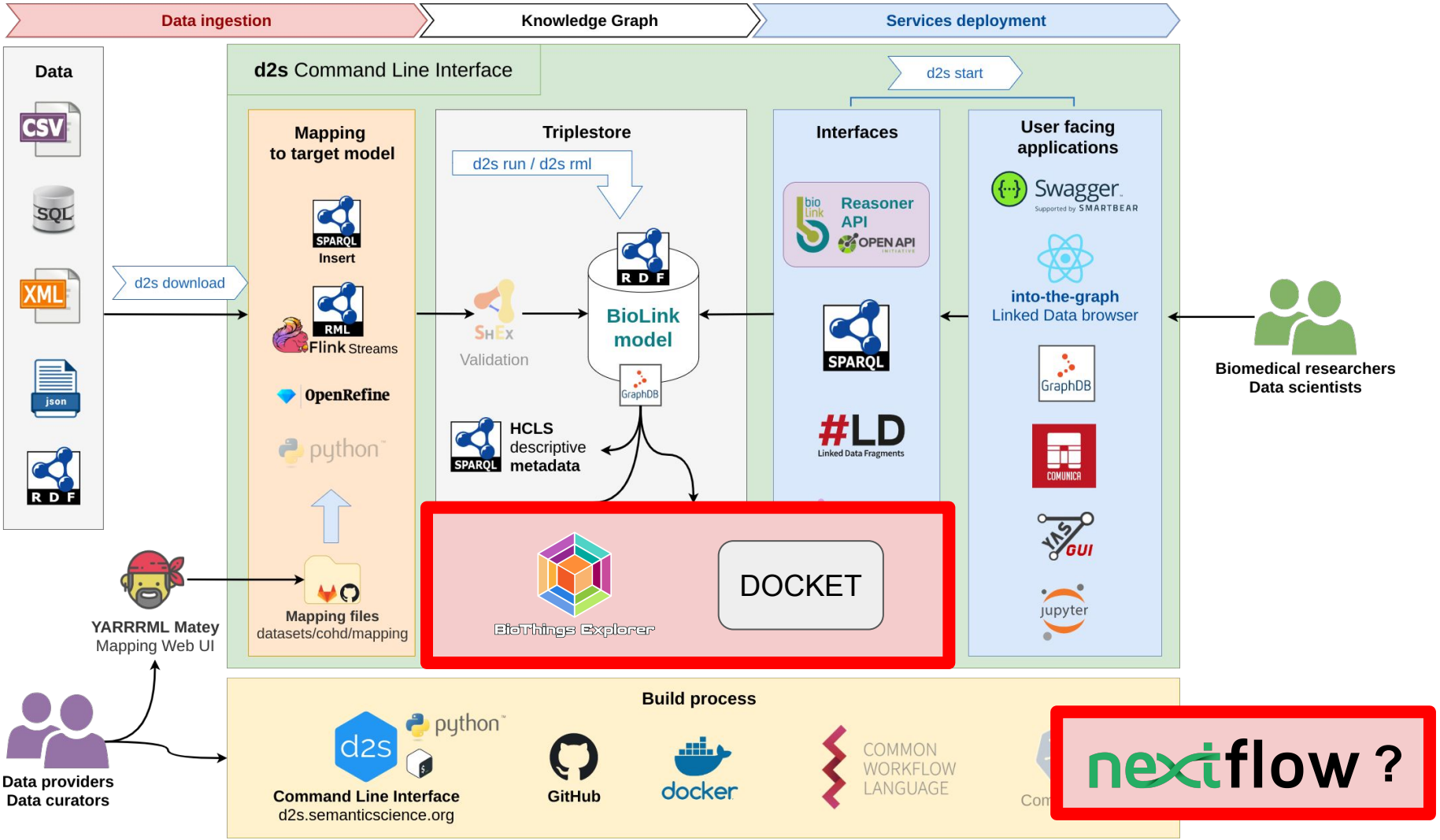
➕ Flexible to add new tools or use custom Notebooks

Use your favorite Translator-compliant “graph database” to store your data



Customize any services deployments using docker-compose

https://d2s.semanticscience.org/docs/d2s-services


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s start
d2s download
Mapping to target model
SPARQL Insert
RML Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
Triplestore
d2s run / d2s rml
SHEX Validation
RDF
BioLink model
GraphDB
SPARQL HCLS descriptive metadata
Interfaces
bio link Reasoner API
OPEN API
SPARQL
#LD Linked Data Fragments
User facing applications
Swagger Supported by SMARTBEAR
into-the-graph Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
YARRRML Matey
Mapping Web UI
BioThings Explorer
DOCKET
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
nextflow ?

Documentation and prototype

Documentation: https://d2s.semanticscience.org

Create a d2s project from the template:
https://github.com/MaastrichtU-IDS/d2s-transform-template

TReK KG and services Web UI (Reasoner API, SPARQL):
http://trek.semanticscience.org

ReasonerStd API at http://api.trek.semanticscience.org

Tool on PyPi: `pip install d2s`

Glad to discuss how d2s can make the deployment of your tool easier!
Or any technical details

The Knowledge Collaboratory

Problem: There is **no easy way for researchers to directly contribute to the body of knowledge that the Translator community relies on**, beyond the resources provided by the KPs. Users remain unable to contribute new facts or even comment on existing facts in a way that could be considered by the reasoner tools or the end users themselves.

Solution: A software infrastructure (databases + api + user interface) to enable collaboration on biomedical knowledge knowledge graphs. The project would emphasize the inclusion of **provenance** (who, where, when and how was it reported), **evidence** (what supports or disputes the assertion?), and **confidence** (how reliable is the statement in terms of evidence?)

- add new hyposeses (“hydroxychloroquine treats COVID”) with evidence from in vitro studies, clinical trials, case reports, obs. health data
- refine and relate hypotheses (“hydroxychloriquine can treat mild COVID patients without cardiovascular risk factors”)
- dispute facts and hypotheses with evidence from scientific studies (or lack thereof)

Users will be guided in **making valid scientific claims with rich context**. They will be able to **retrieve existing knowledge graph** content using Translator infrastructure to have **structured discussions** on the veracity and confidence of assertions, and subsequently propose changes to whether these assertions should be.

nanobench

Publish a new Nanopublication

Publish a new nanopublication below with the chosen template (source) or choose a different template.

Template: Drug Action template

This drug action is a DrugAction .

This drug action has drug http://bio2rdf.org/ add the drug identifier (Bio2RDF prefix:identifier) here .

This drug action has drug action Choose One .

This drug action has phenotype http://bio2rdf.org/ add the phenotype id (Bio2RDF prefix:identifier) here .

This drug action has context " add any specified context for the drug action here " .

This quoted text value " quote the supporting text here " .

This quoted text was quoted from add the URL to the source (e.g. DailyMed URL) of the quoted .

This drug action was derived from add the URL to the source (e.g. DailyMed URL) of the quoted .

This drug action has type of evidence http://purl.obolibrary.org/obo/ECO_ add the evidence code (only the numeric part) from the evider .

Choose One
- Choose One
- is indicated for the treatment of
- is indicated for the management of
- is indicated for the symptomatic relief of
- is indicated as adjunct for
- is contraindicated with
- is associated with risk of

I understand that publishing cannot be undone and that the provided information will be publicly visible and openly connected to my ORCID identifier.

Publish

https://github.com/peta-pico/nanobench

http://purl.org/np/RAfC703YJFOam3pH9xY0s1QYIhDKwZzfRooUnRPu5-xNI

```
sub:da a dao:DrugAction ;
  prov:wasDerivedFrom <https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=452092b4-7a8f-4d19-8113-f1c2a948d3d8> ;
  dao:context "in adults" ;
  dao:drug <http://bio2rdf.org/drugbank:DB01611> ;
  dao:drug-action dao:is-indicated-for-the-treatment-of ;
  dao:evidence <http://purl.obolibrary.org/obo/ECO_0000033> ;
  dao:phenotype <http://bio2rdf.org/snomedct:55464009> .
sub:quote prov:value "Hydroxychloroquine Sulfate Tablets are indicated for the treatment of chronic discoid lupus
erythematosus and systemic lupus erythematosus in adults." ;
  prov:wasQuotedFrom <https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=452092b4-7a8f-4d19-8113-f1c2a948d3d8> .
```

```
sub:assertion prov:wasAttributedTo orcid:0000-0003-4727-9435 .
```

Michel Dumontier (0000-0003-4727-9435) — Thu Apr 23 15:04:16 GMT 2020 — — valid signature

Nanopublication of SLE as an approved indication for Hydroxychloroquine sulfate tables as identified from an FDA structured product labels

Uses DailyMed, Drugbank, SNOMED-CT, and the evidence code ontology.

created and signed by Michel Dumontier (orcid:0000-0003-4727-9435).

Curating COVID treatments

Little is known about which interventions are effective in treating COVID. Preprints and first papers are now coming out.

Goal is to rapidly curate treatment hypotheses and their evidence (or lack thereof)

Define templates, enhance the nanobench tool, and make results available through Translator-compliant interfaces (ReasonerStd API, Smart API)

```
sub:assertion {
  sub:da a dao:DrugAction ;
    prov:wasDerivedFrom sub:quote ;
    dao:drug drugbank:DB00207 , drugbank:DB01611 ;
    dao:drug-action dao:treats ;
    dao:phenotype <http://purl.obolibrary.org/obo/MONDO_0100096> .
  sub:quote prov:value "HYDROXYCHLOROQUINE & AZITHROMYCIN, taken together, have a real chance to
be one of the biggest game changers in the history of medicine. The FDA has moved mountains -
Thank You! Hopefully they will BOTH (H works better with A, International Journal of
Antimicrobial Agents)....." ;
    prov:wasQuotedFrom <https://twitter.com/realDonaldTrump/status/1241367239900778501> .
  <https://twitter.com/realDonaldTrump/status/1241367239900778501> dct:creator
<https://twitter.com/realDonaldTrump> .
}
```